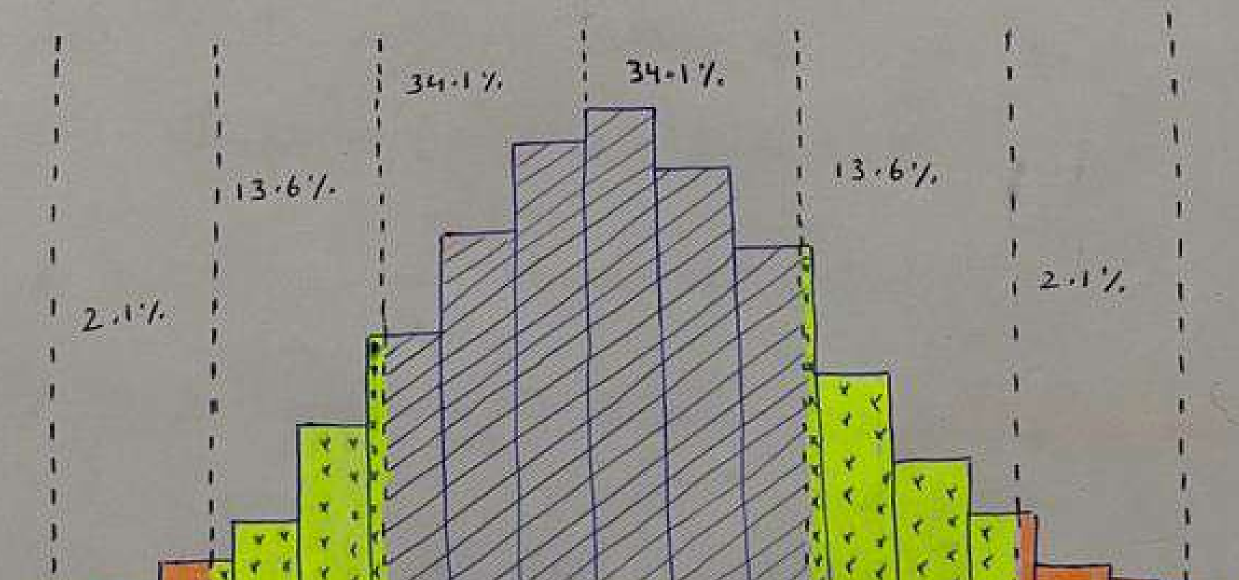
All values ranging b/w lower fench and Higher fench are considered and rest are treated as outliers.

Gaussian / Narmal distribution:

It's continuous prabability distribution that's symmetrical around it's mean and most observations clusters around the central peak. It's describes how the values of a variable are distributed.

- Emperial rule of Narmal distribution:



Emperial rule states that 68% of data fail within range of $(\mu \pm \sigma)$, 95% of data fails within range of $(\mu \pm 2\sigma)$ and 99.7% of data within $(\mu \pm 3\sigma)$ of range.

Standard Normal distribution.

It's also called as Z-distribution. Any normal distribution can be standardized by converting it's values into Z-scores which let's know how many std. deviation (σ) from the mean each values lies. In case of std. Normal distribution value of mean and std. deviation (σ) is 0 and 1.

X = Gaussian Distribution (μ, σ)

Y = St. Narmal Distribution $(\mu = 0, \sigma = 1)$

$$\text{Z-score} = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}} \quad ; \quad \frac{\sigma}{\sqrt{n}} = \text{Std. error.}$$

when we go through each individual element $n = 1$

Thus, Z-score $= \dfrac{x_i - \mu}{\sigma}$

Let, $X = \{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma = 1.41$

Now, calculating Z-score.

$$x_1 = \frac{x_i - \mu}{\sigma} = \frac{1-3}{1.414} = -1.41$$

$$x_2 = \frac{2-3}{1.414} = -0.7$$

$$x_3 = \frac{3-3}{1.414} = 0$$

$$x_4 = \frac{4-3}{1.414} = 0.7$$

$$x_5 = \frac{5-3}{1.414} = 1.414$$

$$Y = \{-1.41, -0.7, 0, 0.7, 1.41\}$$

- In case of standardization:

$\mu = 0 \quad , \sigma = 1$

As per rule of Normal distribution:

99.7% of data lies b/w

$(\mu \pm 3\sigma)$ and putting $\mu = 0, \sigma = 1$

99.7% of data will lie b/w

-3 to $+3$

Normalization:

It's scaling technique method in which data points are shifted and rescaled so that they end up in a range b/w 0 to 1 and is also known as min - max scaling.

Calculation for normalized score:

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

X_{min} and X_{max} are minimum and maxi-
-mum values of the feature/data.

Narmaligation Vs Standardigation.

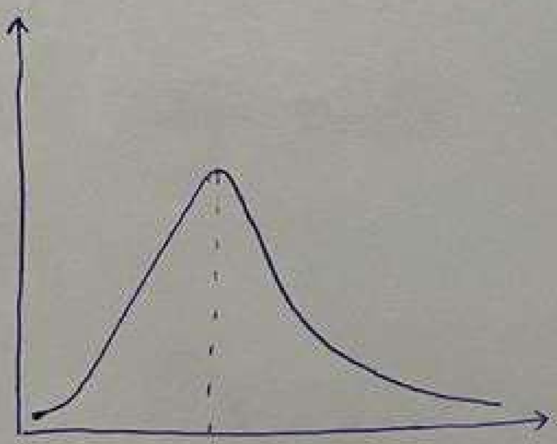
- Normaligation is prefered when data doesn't fallow a Normal distribution. we narmalige value b/w 0 to 1. Useful in such algarithms that do nat assume any distribution of data, like - k nearest neighbor and neutral networks used in deep learning.

- Standardigation is preferred when data fallows Normal distribution. when we can make assumption of data. we have no restrictions to bound data within range as in Normaligation. So good to go when have outliers in data.

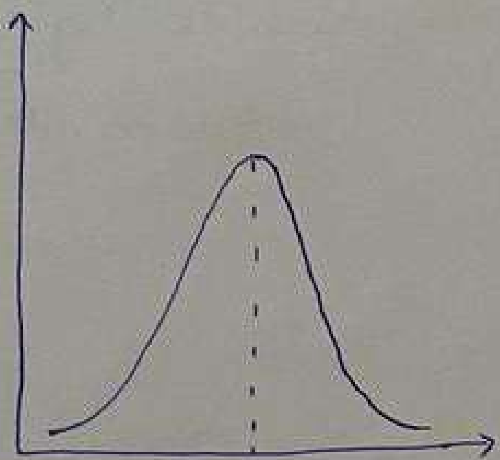
log normal distribution.

Continuous probability distribution of a random variable in which logarithm is Normally distributed.

If random variable X is log normally distributed then $Y = In(X)$ has a normal distribution. Vice-Versa if Y has normal distribution than e^Y ($X = e^Y$) has log normal distribution.



X = log normal distribution



Y = In(x) Normal distribution

Right skewed vs left skewed.

- Distribution is Right / left skewed if, as in the histogram above. Right / left tail (smaller values) is much longer than the left / Right tail (larger values).

In case of Right skewed distribution, bulk of the observations are medium / large with a few observations that are much smaller than the rest.

when distribution is right skewed, mean is often greater than the median.